R E P O R T   R E S U M E S

ED 016 288   EA 001 060

ON THE THEORY AND PROCEDURE FOR CONSTRUCTING A
MINIMAL-LENGTH, AREA-CONSERVING FREQUENCY POLYGON FROM
GROUPED DATA.

BY- CASE, C. MARSTON
NATIONAL CENTER FOR EDUCATIONAL STATISTICS (DHEW)

REPORT NUMBER TN-3   PUB DATE 15 JUL 66
EDRS PRICE MF-$0.25 HC-$0.72 16P.

DESCRIPTORS- *CHARTS, *GRAPHS, DATA ANALYSIS, *THEORIES,
*METHODS, MATHEMATICAL MODELS, GEOMETRY, *STATISTICS,
DISTRICT OF COLUMBIA,

THIS PAPER IS CONCERNED WITH GRAPHIC PRESENTATION AND ANALYSIS OF GROUPED OBSERVATIONS. IT PRESENTS A METHOD AND SUPPORTING THEORY FOR THE CONSTRUCTION OF AN AREA-CONSERVING, MINIMAL LENGTH FREQUENCY POLYGON CORRESPONDING TO A GIVEN HISTOGRAM. TRADITIONALLY, THE CONCEPT OF A FREQUENCY POLYGON CORRESPONDING TO A GIVEN HISTOGRAM HAS REFERRED TO THAT POLYGON FORMED BY CONNECTING THE MIDPOINTS OF THE TOPS OF THE RECTANGLES MAKING UP THE HISTOGRAM. THE MOST IMPORTANT DEFICIENCY IN THE TRADITIONAL FREQUENCY POLYGON IS THAT THE AREA OF ANY SPECIFIC RECTANGLE IN THE UNDERLYING HISTOGRAM IS GENERALLY NOT EQUAL TO THE AREA UNDER THE FREQUENCY POLYGON OVER THE SAME INTERVAL. DUE TO THIS DEFICIENCY, DATA ARE SELDOM PRESENTED IN THE FORM OF A FREQUENCY POLYGON. (HW)

NATIONAL CENTER FOR EDUCATIONAL STATISTICS
Division of Operations Analysis

ON THE THEORY AND PROCEDURE FOR CONSTRUCTING A MINIMAL-LENGTH, AREA-CONSERVING FREQUENCY POLYGON FROM GROUPED DATA

by

C. Marston Case

Technical Note
Number 3

July 15, 1966

OFFICE OF EDUCATION/U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE

NATIONAL CENTER FOR EDUCATIONAL STATISTICS
Alexander M. Mood, Assistant Commissioner

DIVISION OF OPERATIONS ANALYSIS
David S. Stoller, Director

This paper is concerned with the problem of the graphic presentation and analysis of grouped observations. Suppose a set of M observations has been classified on the basis of n contiguous and non-overlapping intervals, *i.e.*, each observation falls into exactly one of the intervals, and then is identified by that interval. Observations so classified and identified are said to be grouped. The resulting recorded data is then in the form of a set of n+1 interval boundaries and a set of n integers (N_1, N_2, . . . N_n), where N_i indicates the number of observations in the i^{th} interval and $\sum N_i=M$.

Such observations are often graphically presented in a histogram consisting of n rectangles (contiguous and nonoverlapping) with the widths (w_i) of the rectangles proportional to the lengths of the grouping intervals and the heights (h_i) of the rectangles proportional to N_i/w_i. The products $h_i w_i$ are consequently proportional to the corresponding percentage of observations in the i^{th} interval, *i.e.*, there exists a constant c such that $ch_i w_i=N_i/M$. The constant c is generally incorporated into the scale used for drawing the histogram and will be omitted henceforth.

If the last (n^{th}) interval is not finite the proportion in it is not represented in the histogram. In this case $\sum^{n-1} h_i w_i = 1 - N_n/M$; otherwise $\sum h_i w_i = 1$.

Traditionally the concept of a frequency polygon (FP) corresponding to a given histogram has meant that polygon formed by connecting the midpoints of the tops of the rectangles making up the histogram. (Ref. Hald 49-51, Dixon & Massey 8-9) The most important deficiency in the traditional frequency polygon is that the area of any specific rectangle in the underlying histogram is generally not equal to the area under the frequency polygon over the same interval. Due to this deficiency data are seldom presented in the form of a frequency polygon. The purpose of this paper is to present a method and supporting theory for the construction of an area-conserving, minimal length frequency polygon corresponding to a given histogram.

For purposes of this paper histograms and frequency polygons will be considered to have the following four parts (see figure 1):

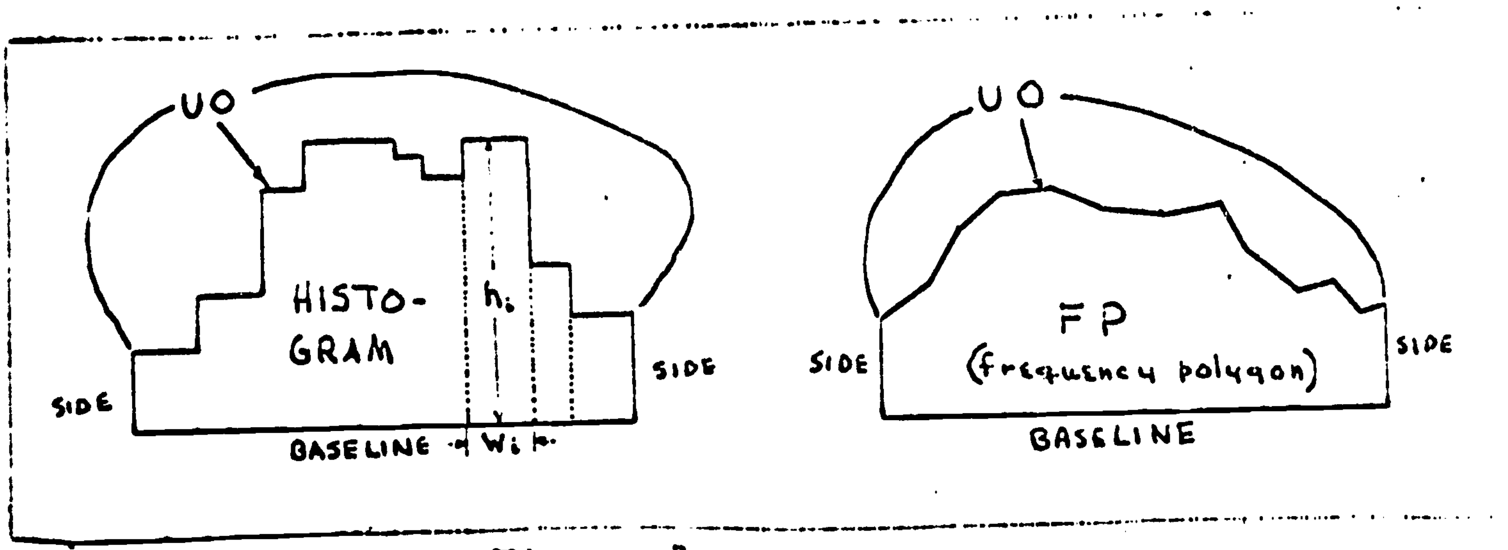

Figure 1

1) the base line, the horizontal line of length $\sum w_i$ upon which the histogram or FP is constructed;

2) & 3) the two sides, which rise vertically from both ends of the base line;

4) the upper outline (UO) which comprises the remainder of the histogram or FP, a function consisting of connected line segments which connects the two sides across the top.

The base line will be assumed to be that portion of an axis of abcissas from zero to $\sum w_i$.

In terms of this nomenclature, the method this paper presents is that of the construction of an FP corresponding to a given n-interval histogram such that

1) letting $z_i = \sum_{k=1}^{i} w_k$,

$$\int_0^{z_i} \text{FPUO } dx = \sum_{k=1}^{i} w_k h_k \text{ , } i=1, \ldots n$$

(or i=1, . . . n-1 if the last interval is not finite);

2) the FPUO is the minimal length FPUO consisting of 2n connected line segments such that there are no more than two line segments over any given interval.

We will first show which two-segmented UO's conserve the area of a single rectangle (see Figure 2). We are given the rectangle ABCD, its sides being segments of vertical lines L_1 and L_2. An UO is to be constructed consisting of the two line segments EP and PF with E on L_1 a distance q below B and with F on L_2 a distance r above c.

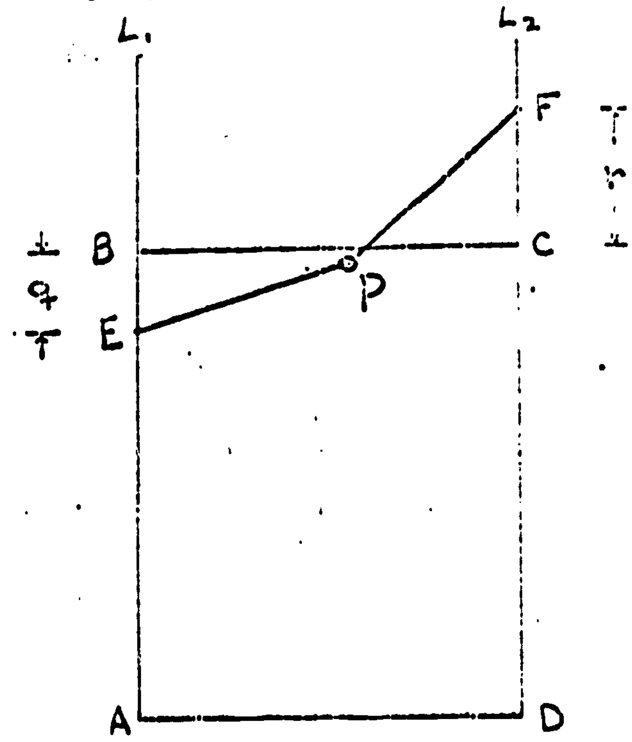

Figure 2

In order to conserve the area of ABCD we restrict q and r as follows: $q \leq AB$, $r \geq -CD$. For such an arbitrary point (q,r) in the q-r plane we seek the locus of points P between L_1 and L_2 such that the area of the pentagon AEPFD is the same as that of the rectangle ABCD. Locate the point O which is the midpoint of the line segment BC (Figure 3). Extend the line FO to intersect L_1 at E' and the line EO to intersect L_2 at F'. Draw the line E'F'.

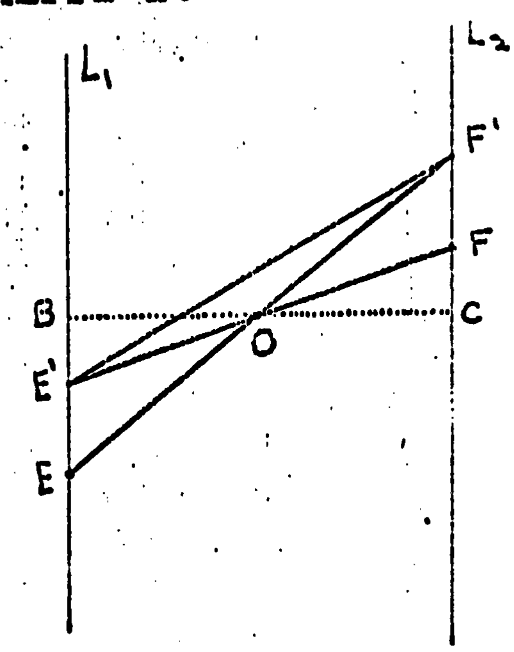

Figure 3

Theorem I

E'F' is the set of points such that for any point P on E'F', area AEPFD = area ABCD.

Proof

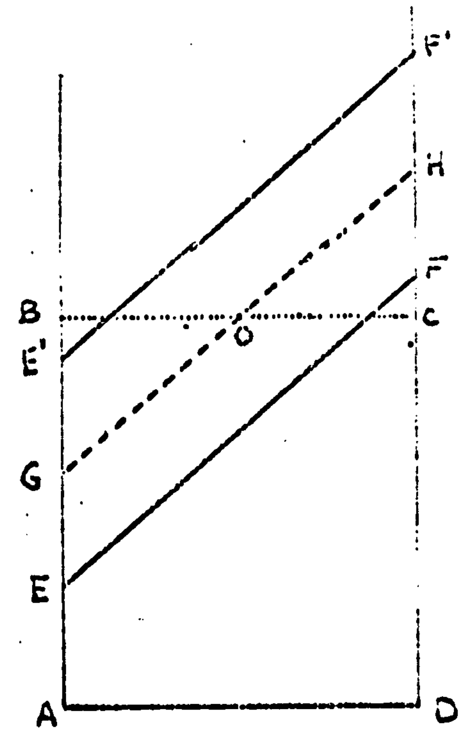

Figure 4

Draw in the auxilliary lines EF and GH where GH is parallel to E'F' and passes through O (figure 4). The area of the trapezoid AGHD is the same as the area of ABCD since $\triangle GOB = \triangle COH$. Trapezoid AGHD = trapezoid AEFD + parallelogram EGHF and parallelogram EGHF is $\frac{1}{2}$ parallelogram EE'F'F. It remains to show that any triangle EP'F with P' on E'F' has half of the area of parallelogram EE'F'F. Draw PP' parallel to FF' (figure 5). Then $\triangle$ EPP'= $\frac{1}{2}$ parallelogram EE'PP' and $\triangle$ FPP' = $\frac{1}{2}$ parallelogram EE'F'F. $\triangle EPF = \triangle EPP' + \triangle FPP' = \frac{1}{2}$ parallelogram EE'F'F. Q.E.D.

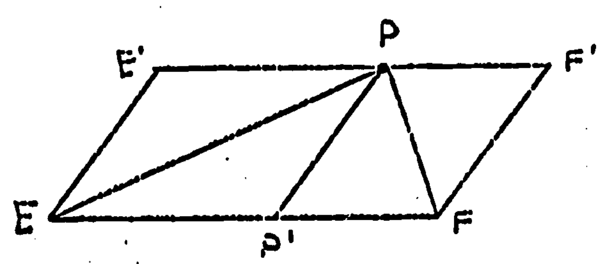

Figure 5

We also need to know which P minimizes the sum of the lengths of line segments EP+PF.

Theorem II

Given two points E and F and a line L parallel to the line through E and F, the point P on L which minimizes EP + PF is at the intersection of L and the perpendicular bisector of the segment EF (figure 6).

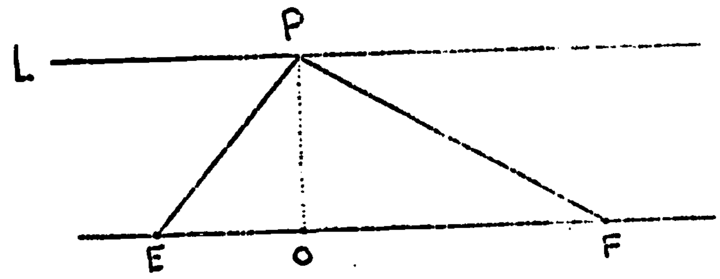

Figure 6

Proof

Let O be a point between E and F and d the length of segment EF. Then for some q, $0 \leq q \leq 1$, EO=qd and FO=(1-q)d. If PO=c, then

$$EP + PF = \sqrt{q^2d^2+c^2} + \sqrt{(1-q)^2d^2+c^2} .$$

The derivative of this with respect to q is

$$\frac{2qd^2}{\sqrt{q^2d^2+c^2}} - \frac{2(1-q)d^2}{\sqrt{(1-q)^2d^2+c^2}}$$

This is made equal to zero by letting $q=\frac{1}{2}$, which means that EP+PF is minimized when EO=FO, that is, when PO is the perpendicular bisector of EF.

Theorem III

If ABCD is a rectangle formed between L_1 and L_2 and E, F, E', F' are points determined by the quantities q and r as indicated above and P is at the intersection of the perpendicular bisector of EF and E'F', then the length of EP (and FP) is

$$\sqrt{\frac{w^2(q-r)^2}{w^2+(q+r)^2} + \frac{w^2+(q+r)^2}{4}}$$

Proof

Let a be the distance between the lines E'F' and EF and let 2b be the length of EF and E'F'. (figure 7) Construct FG parallel to BC. $(2b)^2 = w^2+(q+r)^2$

$$b^2 = \frac{w^2+(q+r)^2}{4}$$

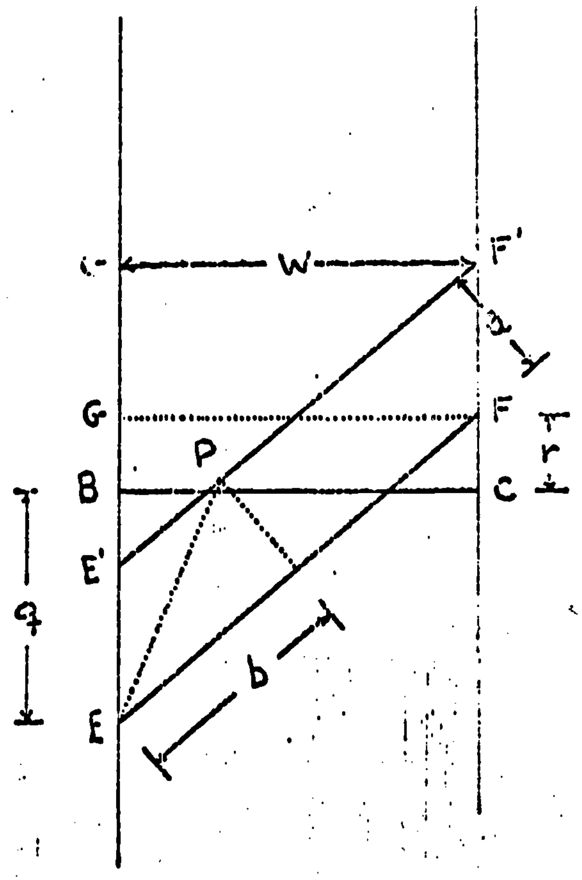

Figure 7

Construct (figure 8) E'H perpendicular to EF, the length of E'H=a.

EE'=BE-FC=q-r

Q is the midpoint of EF

QQ' is perpendicular to BE and QQ' = $\frac{1}{2}$w. Thus since $\triangle$ EE'H is similar to $\triangle$ EQQ'

$\frac{w}{2}$: b as a : q-r

or

$$\frac{w}{2b} = \frac{a}{q-r}$$

or

$$a = \frac{w(q-r)}{2b}$$

$$a^2 = \frac{w^2(q-r)^2}{4} \cdot \frac{4}{w^2+(q+r)^2}$$

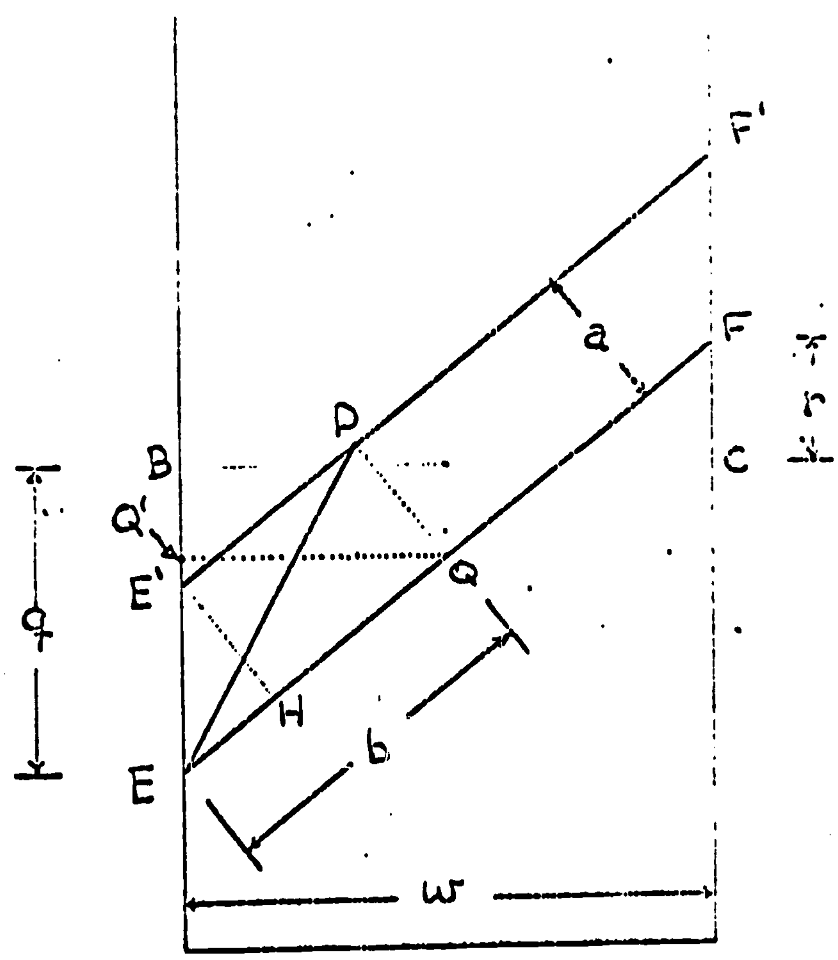

FIGURE 8

Thus

$$(EP)^2 = a^2+b^2 = \frac{w^2(q-r)^2}{w^2+(q+r)^2} + \frac{w^2 + (q+r)^2}{4}$$

Q.E.D.

For future reference let us identify this UO length thus:

$$g(q,r,w) = 2EP$$

$$= \sqrt{\frac{4w^2(q-r)^2}{w^2+(q+r)^2} + w^2 + (q+r)^2}$$

In review, Theorems I and II show how to find the minimal length area-conserving 2-segment UO for a given rectangle of width w and given values of q and r. Theorem III establishes the length of this UO. If q and w are given and we seek that r which gives us the minimal length UO we find the derivative,

$$\frac{\partial g(q,r,w)}{\partial r} = \frac{1}{2g}\left[\frac{2w(q-r)}{w+(q+r)^2} - \frac{2w(q+r)(q-r)^2}{\left[w+(q+r)^2\right]^2} + 2(q+r)\right]$$

Consider w as fixed and the above derivative to be a surface with reference to the q-r plane. Consider the points (the function) on the q-r plane at which this derivative surface passes through the plane; *i.e.*, consider the points at which the surface is zero. A computer program was written which has shown this function to be monotonic increasing for q in the range of our application. A computer subroutine, named UZBEK, using an iterative procedure (internal halving) has been written to find the required r for a given q, w (and an r-minimum to prevent the UO from passing below the base line of the rectangle).

We now consider the problem of constructing the minimal length, area-conserving 4-segmented UO over a given 2-rectangle histogram with only the sides of the FP given. Let q_1 be the left side of the FP

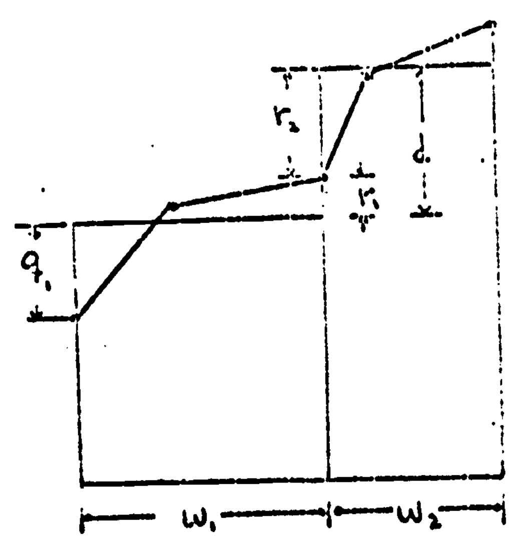

Figure 9

histogram minus the left side of the FP and q_2 be the right side of the FP minus the right side of the histogram (see figure 9). Let the difference between the height of the right rectangle and the left rectangle be $r_1+r_2=d$. We now seek the pair of numbers (r_1, r_2) such that $g(q_1, r_1, w_1) + g(q_2, r_2, w_2)$ is minimized. We again consider the derivative surface $\partial g/\partial r$ over the q-r plane for a fixed w. Let $f_1(r) = \partial g(q_1, r, w_1)/\partial r$ and $f_2(r) = \partial g(q_2, r, w_2)/\partial r$. These are the intersections of the planes $q=q_1$ and $q=q_2$ with the w_1 type derivative surface and the w_2 type respectively. Notice that since $r_1=d-r_2$, a change in r_1 produces a change in r_2 which differs from that of r_1 only in sign. Thus, when the point marking the division of d into r_1 and r_2 is moved along the left side of the right rectangle the rate of change of the length of the left portion of the FPUO is opposite in sign to that of the right portion. We want the overall derivative, or the sum of these two derivatives, to be zero. This is accomplished when

$f_1(r_1) = f_2(r_2)$ and $r_1 + r_2 = d$. This is essentially a Lagrangian multiplier problem but requires numerical methods. Another computer subroutine, named ORYX, has been written to provide the minimal length, area-conserving 4-segmented FPUO given q_1, q_2, d, w_1, and w_2.

We are now ready to describe a procedure for constructing an FP for the general histogram made of n rectangles. In the histogram UO there are n-1 vertical line segments which we will call "risers". The midpoints of the risers are used for the first stage estimates of the height of the required FP at the interval boundaries. The first such riser midpoint is used to obtain an estimate of the height of the left side of the FP (using UZBEK). This left side point, along with the midpoint of the second riser, is used to obtain a new point on the first riser (using ORYX). Then the new first riser point and the third riser midpoint are used to obtain a new second riser point. This method is carried out across the histogram until only the last riser midpoint remains unchanged. Then (using UZBEK) the last riser midpoint is used to estimate the right side. Finally the right side point and the new point on the next to the last riser are used to obtain a new point on the last riser. This procedure is repeated across the histogram several times until the point with the greatest change in its position on its riser is less than some prespecified number. The change tested is expressed as the fraction of the riser traversed during the pass. This procedure brings the maximum change under .0001 in about n passes.

To complete the required frequency polygon, we need the points (P) within the intervals which are derived from the points on the risers. Let the origin for derivation of a given P be the top left corner of the histogram rectangle for that interval. (See figure 10) The point P is at the intersection of

1) the line through $(0,-r)$ with slope $\frac{q+r}{w}$ and

2) the line through $\left(\frac{w}{2}, \frac{q-r}{2}\right)$ with slope $-\frac{w}{q+r}$.

Solving the resulting simultaneous equations in x and y we obtain

$$x = \frac{w}{2} + \frac{w(r^2-q^2)}{(q+r)^2+w^2} \quad \text{and}$$

$$y = \frac{(q-r)\left[w^2-(q+r)^2\right]}{2\left[w^2+(q+r)^2\right]}$$

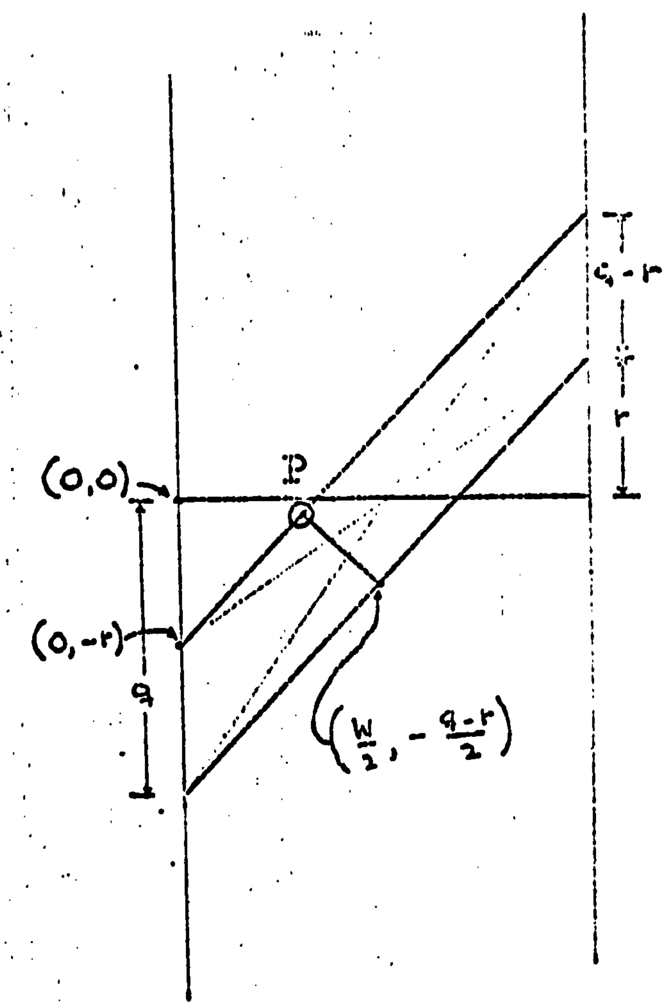

Figure 10

This completes Part I of the outline of the logic and method of constructing the required frequency polygon. An important characteristic of such FPs is that their shapes are not invariant when the width (w_i) scale is changed. Since this scale is essentially arbitrary (we can measure lengths in inches or meters or furlongs, etc.), we need a criterion which establishes a standard scale for a given set of grouped data. This topic will be taken up in the next part. Other topics to be considered are:

1) uses of the minimal length area-conserving frequency polygon, such as:

   a) comparison of hypothesized theoretical distributions and corresponding actual observations

   b) interpolation of percentiles derived from grouped data

2) sample graphs of frequency polygons.

REFERENCES

Dixon, W.J. and F.J. Massey, Jr. Introduction to Statistical Analysis, New York: McGraw Hill, 1957.

Hald, A. Statistical Theory with Engineering Applications, New York : John Wiley, 1952.